文化と言語資源

著者	田中 穂積
雑誌名	日本語科学
巻	10
ページ	3-3
発行年	2001-10-30
URL	http://id.nii.ac.jp/1328/00002066/

文化と言語資源

田中　穂積

通時的にであれ，共時的にであれ，大量の例文を集めておき，それらを網羅的に分析することが言語の研究や辞書作りにとってきわめて重要であることは言を待たない。

日本語ワープロがまだ研究段階であった1970年代半ばには，電子化された文，すなわち磁気テープなどの電子媒体上に書き込まれた日本語文の量は，現在とは比べものにならないほど少量であった。丁度このころ，筆者が所属していた研究所に，古文の用例を書き込んだ大量のカードをめくりながら，単語の使用頻度を調べていたアルバイトの学生がいた。用例カードを電子化しておけば，用例カードをめくるという単純作業をコンピュータに代行させることができるはずであるが，彼は「いや，このカードの厚さが，私がこれまでどれほど単純作業に耐えて努力をしてきたかをはかる尺度になります。カードの厚さが卒業研究の合否をきめるんです。コンピュータがこの単純作業を代行してしまったら困るんですよ」と笑いながら話してくれたことを思い出す。当時は使用頻度に基づく語彙調査そのものが研究になり得た時代であった。

話したり書いたりした文を大量に集めたものをコーパスとよぶ。辞書や各種コーパス，言葉を理解するために必要な知識の体系などを電子化したものを総称して「言語資源」とよぶ。言語の研究や辞書の構築以外にも「言語資源」を利用した研究が最近盛んである。音声認識システムの認識精度向上に「言語資源」が大きな役割を果たしたことは良く知られている。最近の音声認識システムでは，大量の「言語資源」から，各単語の直後に現れる単語の頻度情報（統計データ）をあらかじめ獲得しておき，この統計データを用いて，認識結果の候補にもっともらしさの順位を付けて音声認識の精度をあげている。自然言語処理の分野では文の係り受け関係の解析は重要であるが，係る側の単語と係り先の単語についての統計データがあれば，それを文の係り受け解析に利用することができる。そのためには，人手を介して正しい係り受け関係を付与した多数の文を用意しておかなければならない。

最近では，コンピュータを用いて，辞書の説明文から単語間の関係（「蝉」の項目に「昆虫の一種」とあれば，「昆虫」は「蝉」の上位に属すなどという関係）を自動的に抽出したり，翻訳結果を並記したコーパスがあれば，そこから翻訳用の知識を抽出したり，多数の文書を自動的に分類する研究が活発である。この種の研究をテキストマイニングとよぶ。このとき係り受け解析やテキストマイニングのもとになる「言語資源」の量は多ければ多いほど良いのであるが，「言語資源」の整備には時間と労力とお金がかかる。そこで，構築した「言語資源」を共有し合うことが望まれる。米国，欧州，そしてつい最近韓国でも，「言語資源」の整備とそれを多くの人に安価に流通させる機構を政府の援助で立ちあげている。わが国の現状はどうか。一刻の猶予もならないという危機感にかられて，筆者は２年前から言語資源共有機構（GSK）の設立を政府関係者にはたらきかけているが理解がえられない。GSK の活動はボランティアベースで細々と続けているというのが現状である。一朝一夕にはできない「言語資源」の整備に国からの援助がないのは，文化国家日本としていささか恥ずかしいことだと思う。言葉を大切にするかどうかは，その国の文化の程度を示すバロメータだと思うからである。